Agenda

- [10] Review data logs
- [10] Getting data
- [10] Grad student presentation on open data papers
- [20] Cleaning data
- [10] Presentation crit
- [5] Homework prep
data log
pair & share

the most nefarious?
the most benign?
the most surprising?
Getting data
Sources of Data

- Official sources (ie. govt agency)
- Advocacy / interest groups
- Personal knowledge
- Make it yourself
If the data doesn't exist?

https://syriancivilwarmap.com/

https://icwatch.wikileaks.org/

Left © Syrian Civil War Map; right © ICWATCH. All rights reserved. This content is excluded from our Creative Commons license. For more information, see http://ocw.mit.edu/help/fac-fair-use/
Document Object Model (ie. "the DOM")
Tools to Scrape

- **BeautifulSoup**
- **mechanize**
- **import.io**
- **jQuery in the browser**
- **requests**

**Axes:**
- x-axis: easy to learn vs. hard to learn
- y-axis: does lots of things vs. does one thing

**Points:**
- **Chrome scraper extension:** easy to learn, does one thing
- **import.io:** does lots of things, hard to learn
- **requests:** hard to learn, does lots of things

How have you seen data stored?
Storage strategies

- .csv files
- relational databases
- non-relational databases
- text files
- .pdf files
- HTML tables
What is "clean" data?
Clean Data

- **Consistency**: are observations always entered the same?
- **Completeness**: do you have coverage of the topic?
- **Usability**: machine readability?
- **Atomicity**: row-based normalization

Since we use machines to operate on data, machine-readability is a strong criteria.

And don't forget about the metadata!

See the Quartz guide to bad data
About
"Tidy" Data

Tools to clean

Excel

- Open Refine
- Trifacta
- Wrangler

Textract

messytables

PDFtables

does lots of things
does one thing

easy to learn

find & replace

Tabula
geocod.io
BatchGeo

regular expressions

hard to learn
Cleaning geographic data

**Geoparsing**: finding references to geographic places in text

tricky, but my [Cliff tool](#) does some of this

**Geocoding**: turning an address into latitude/longitude coordinates

[BatchGeo](#) can do a lot for you for free
Getting data out of PDF files

assuming the text is readable:

Let's open up an example PDF and try out Tabula (the best I've seen so far)

If you're a programmer, pdftables is a useful option

if it is an image:

you're in trouble - the automated OCR toolchain isn't great
Cleaning text / numbers

misspellings? try OpenRefine to cluster them

extracting data? try regular expressions (use a cheatsheet) (learn it yourself)

splitting columns? remember Excel can do some of this

anonymizing? scrubadub.io is an in-progress tool to help
Using image data

You can analyze images qualitatively and quantitatively by repurposing tools like Google Photos.

© Google. All rights reserved. This content is excluded from our Creative Commons license. For more information, see http://ocw.mit.edu/help/faq-fair-use/
Another critique
A more complex example

U.S. Daily Temperature Anomalies 1964-2013

What is an anomaly?

Every day, the Global Historical Climatology Network collects temperatures from 90,000 weather stations. Dating back as far as the late 1700s, the records provide an incredible source of insight into our changing climate.

These criteria enabled us to track the last 50 years of temperature anomalies and categorize them into four types.

COLD anomalies occur on days when the daily high or low temperature falls below
homework

● install Tableau

● read stuff

● grad student to present reading on machine learning & big data